```r
library(caret)
library(dplyr)
library(scatterplot3d)
# Load the data
data("GermanCredit")
```

Perform cluster analysis of the data for market segmentation.

Overview of the data

```r
head(GermanCredit)
```

```
##   Duration Amount InstallmentRatePercentage ResidenceDuration Age
## 1        6   1169                         4                 4  67
## 2       48   5951                         2                 2  22
## 3       12   2096                         2                 3  49
## 4       42   7882                         2                 4  45
## 5       24   4870                         3                 4  53
## 6       36   9055                         2                 4  35
##   NumberExistingCredits NumberPeopleMaintenance Telephone ForeignWorker
## 1                     2                       1         0             1
## 2                     1                       1         1             1
## 3                     1                       2         1             1
## 4                     1                       2         1             1
## 5                     2                       2         1             1
## 6                     1                       2         0             1
##   Class CheckingAccountStatus.lt.0 CheckingAccountStatus.0.to.200
## 1  Good                          1                              0
## 2   Bad                          0                              1
## 3  Good                          0                              0
## 4  Good                          1                              0
## 5   Bad                          1                              0
## 6  Good                          0                              0
##   CheckingAccountStatus.gt.200 CheckingAccountStatus.none
## 1                            0                          0
## 2                            0                          0
## 3                            0                          1
## 4                            0                          0
## 5                            0                          0
## 6                            0                          1
##   CreditHistory.NoCredit.AllPaid CreditHistory.ThisBank.AllPaid
## 1                              0                              0
## 2                              0                              0
## 3                              0                              0
## 4                              0                              0
## 5                              0                              0
## 6                              0                              0
##   CreditHistory.PaidDuly CreditHistory.Delay CreditHistory.Critical
## 1                      0                   0                      1
## 2                      1                   0                      0
## 3                      0                   0                      1
## 4                      1                   0                      0
## 5                      0                   1                      0
## 6                      1                   0                      0
##   Purpose.NewCar Purpose.UsedCar Purpose.Furniture.Equipment
## 1              0               0                           0
## 2              0               0                           0
## 3              0               0                           0
## 4              0               0                           1
## 5              1               0                           0
## 6              0               0                           0
##   Purpose.Radio.Television Purpose.DomesticAppliance Purpose.Repairs
## 1                        1                         0               0
## 2                        1                         0               0
## 3                        0                         0               0
## 4                        0                         0               0
## 5                        0                         0               0
## 6                        0                         0               0
##   Purpose.Education Purpose.Vacation Purpose.Retraining Purpose.Business
## 1                 0                0                  0                0
## 2                 0                0                  0                0
## 3                 1                0                  0                0
## 4                 0                0                  0                0
## 5                 0                0                  0                0
## 6                 1                0                  0                0
##   Purpose.Other SavingsAccountBonds.lt.100 SavingsAccountBonds.100.to.500
## 1             0                          0                              0
## 2             0                          1                              0
## 3             0                          1                              0
## 4             0                          1                              0
## 5             0                          1                              0
## 6             0                          0                              0
##   SavingsAccountBonds.500.to.1000 SavingsAccountBonds.gt.1000
## 1                               0                           0
## 2                               0                           0
## 3                               0                           0
## 4                               0                           0
## 5                               0                           0
## 6                               0                           0
##   SavingsAccountBonds.Unknown EmploymentDuration.lt.1
## 1                           1                       0
## 2                           0                       0
## 3                           0                       0
## 4                           0                       0
## 5                           0                       0
## 6                           1                       0
##   EmploymentDuration.1.to.4 EmploymentDuration.4.to.7
## 1                         0                         0
## 2                         1                         0
## 3                         0                         1
## 4                         0                         1
## 5                         1                         0
## 6                         1                         0
##   EmploymentDuration.gt.7 EmploymentDuration.Unemployed
## 1                       1                             0
## 2                       0                             0
## 3                       0                             0
## 4                       0                             0
## 5                       0                             0
## 6                       0                             0
##   Personal.Male.Divorced.Seperated Personal.Female.NotSingle
## 1                                0                         0
## 2                                0                         1
## 3                                0                         0
## 4                                0                         0
## 5                                0                         0
## 6                                0                         0
##   Personal.Male.Single Personal.Male.Married.Widowed
## 1                    1                             0
## 2                    0                             0
## 3                    1                             0
## 4                    1                             0
## 5                    1                             0
## 6                    1                             0
##   Personal.Female.Single OtherDebtorsGuarantors.None
## 1                      0                           1
## 2                      0                           1
## 3                      0                           1
## 4                      0                           0
## 5                      0                           1
## 6                      0                           1
##   OtherDebtorsGuarantors.CoApplicant OtherDebtorsGuarantors.Guarantor
## 1                                  0                                0
## 2                                  0                                0
## 3                                  0                                0
## 4                                  0                                1
## 5                                  0                                0
## 6                                  0                                0
##   Property.RealEstate Property.Insurance Property.CarOther
## 1                   1                  0                 0
## 2                   1                  0                 0
## 3                   1                  0                 0
## 4                   0                  1                 0
## 5                   0                  0                 0
## 6                   0                  0                 0
##   Property.Unknown OtherInstallmentPlans.Bank OtherInstallmentPlans.Stores
## 1                0                          0                            0
## 2                0                          0                            0
## 3                0                          0                            0
## 4                0                          0                            0
## 5                1                          0                            0
## 6                1                          0                            0
##   OtherInstallmentPlans.None Housing.Rent Housing.Own Housing.ForFree
## 1                          1            0           1               0
## 2                          1            0           1               0
## 3                          1            0           1               0
## 4                          1            0           0               1
## 5                          1            0           0               1
## 6                          1            0           0               1
##   Job.UnemployedUnskilled Job.UnskilledResident Job.SkilledEmployee
## 1                       0                     0                   1
## 2                       0                     0                   1
## 3                       0                     1                   0
## 4                       0                     0                   1
## 5                       0                     0                   1
## 6                       0                     1                   0
##   Job.Management.SelfEmp.HighlyQualified
## 1                                      0
## 2                                      0
## 3                                      0
## 4                                      0
## 5                                      0
## 6                                      0
```

```
data <- GermanCredit[, c("Duration", "Amount", "InstallmentRatePercentage", "ResidenceDuration", "Age",
```

I've selected the following numeric variables

```
colnames(data)
```

```
## [1] "Duration"                  "Amount"
## [3] "InstallmentRatePercentage" "ResidenceDuration"
## [5] "Age"                       "NumberExistingCredits"
## [7] "NumberPeopleMaintenance"
```

Scale the data

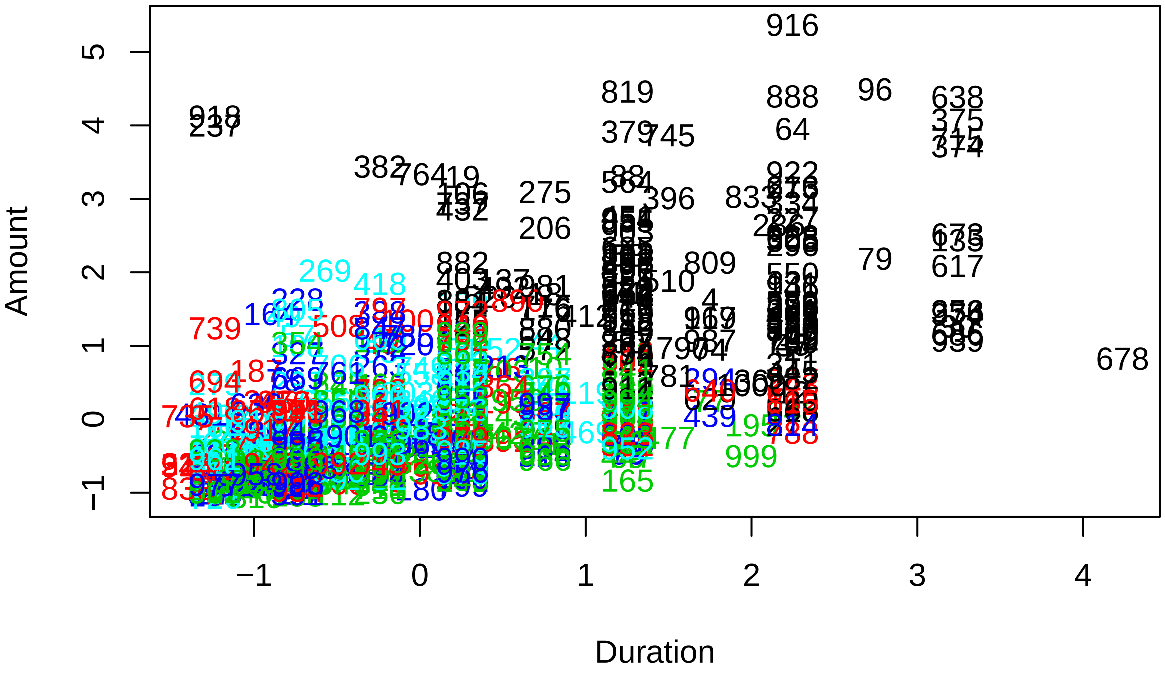
```
standard.data <- scale(data)
```

kmeans

```
res_km <- kmeans(standard.data, centers = 5, nstart = 100)
print(table(res_km$cluster))
```

```
##
##   1   2   3   4   5
## 139 139 337 183 202
```

```
plot(standard.data[,1],standard.data[,2],type="n",xlab="Duration",ylab="Amount")
text(standard.data[,1],standard.data[,2],labels=1:nrow(standard.data),col=res_km$cluster)
```



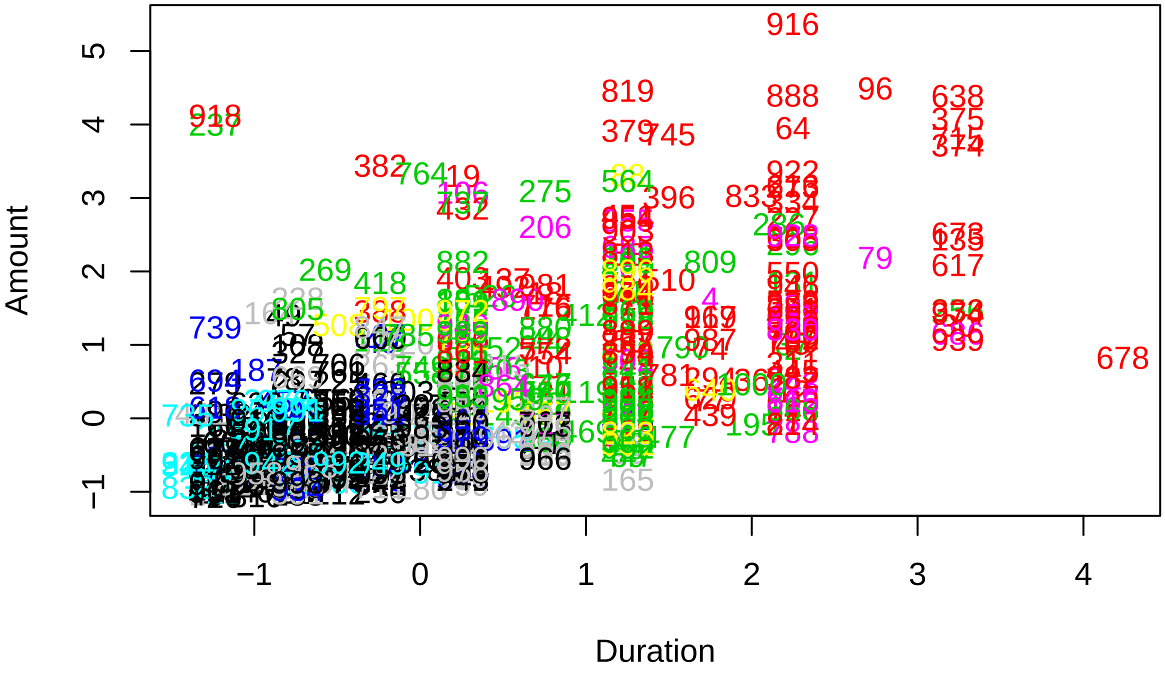
##

komeans

```r
source('komeans.R')
komean_res <- komeans(standard.data,nclust=5,lnorm=2,tolerance=.001,nloops = 120,seed=3)
```

Plot the clusters

```r
plot(standard.data[,1],standard.data[,2],type="n",xlab="Duration",ylab="Amount")
text(standard.data[,1],standard.data[,2],labels=1:nrow(standard.data),col=komean_res$Group)
```



3. Extract 2-10 k-means clusters using the variable set. Present the Variance-Accounted-For (VAF or R-square). the local optima problem is big for all the clustering and latent class methods. So I run it 50-100 random starts.

```r
set.seed(123)
# split the data
train <- sample(1:nrow(standard.data), size = 0.7 * nrow(standard.data))

VAFS_train <- numeric(10)
VAFS_holdout <- numeric(10)
km_train_results <- list()
km_hold_results <- list()
for(i in 2:10){
  # train
  km_res <- kmeans(standard.data[train,], centers = i, nstart = 50 )
  km_train_results[[i]] <- km_res
  VAF <- km_res$betweenss/ km_res$totss
  VAFS_train[[i]] <- VAF
  # holdout
  km_res <- kmeans(standard.data[-train,], centers = km_res$centers, nstart = 50 )
  km_hold_results[[i]] <- km_res
  VAF <- km_res$betweenss/ km_res$totss
  VAFS_holdout[[i]] <- VAF

}
res <- data.frame(k = 2:10, VAF_train = VAFS_train[2:10],VAF_hldout = VAFS_holdout[2:10])
res
```

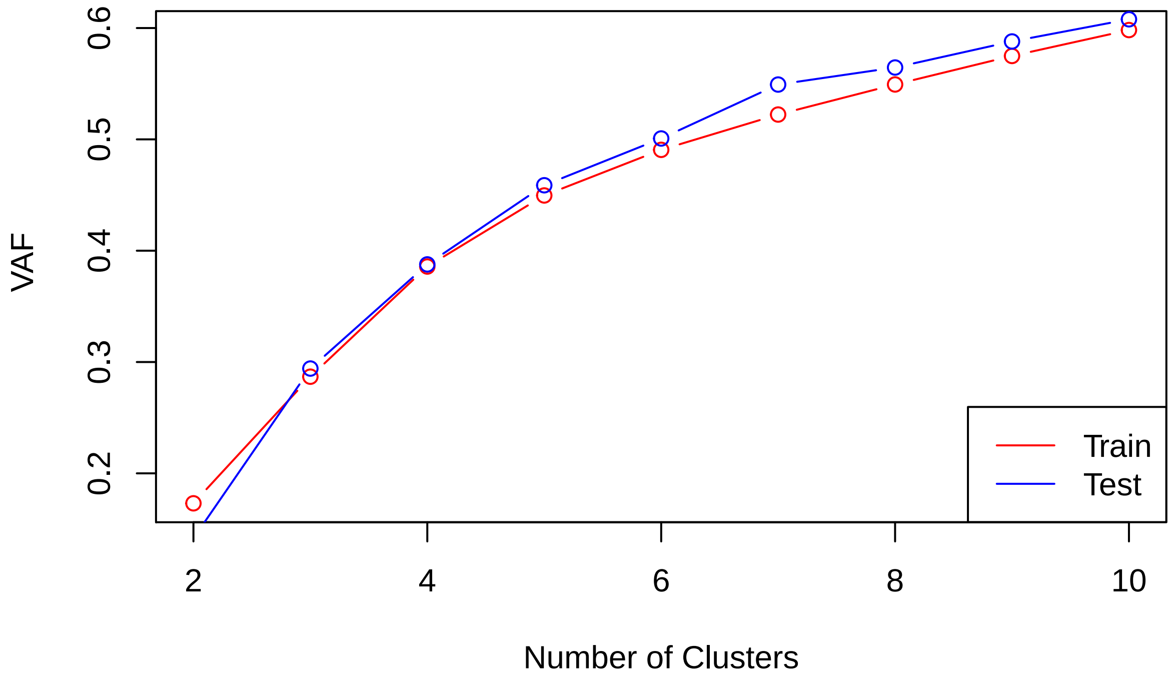
```
##    k VAF_train VAF_hldout
## 1  2 0.1730995  0.1416352
## 2  3 0.2868387  0.2941430
## 3  4 0.3857962  0.3876859
## 4  5 0.4496079  0.4587734
## 5  6 0.4905909  0.5008052
## 6  7 0.5222883  0.5492414
## 7  8 0.5493096  0.5645464
## 8  9 0.5749203  0.5879431
## 9 10 0.5981831  0.6078495
```

4. Scree tests to choose appropriate number of k-means clusters

```
From the scree plot 5 cluster seems suitable. The elbow point is at 5 clusters.
```

5. Show the scree plot.

```r
plot(res$k,  res$VAF_train, type = 'b',col='red', xlab = "Number of Clusters",
     ylab = " VAF")
lines(res$k,  res$VAF_hldout, type = 'b',col='blue', xlab = "Number of Clusters")
legend(x = "bottomright",
       legend = c("Train","Test"),
       lty = c(1,1),
       col = c("red","blue"))
```



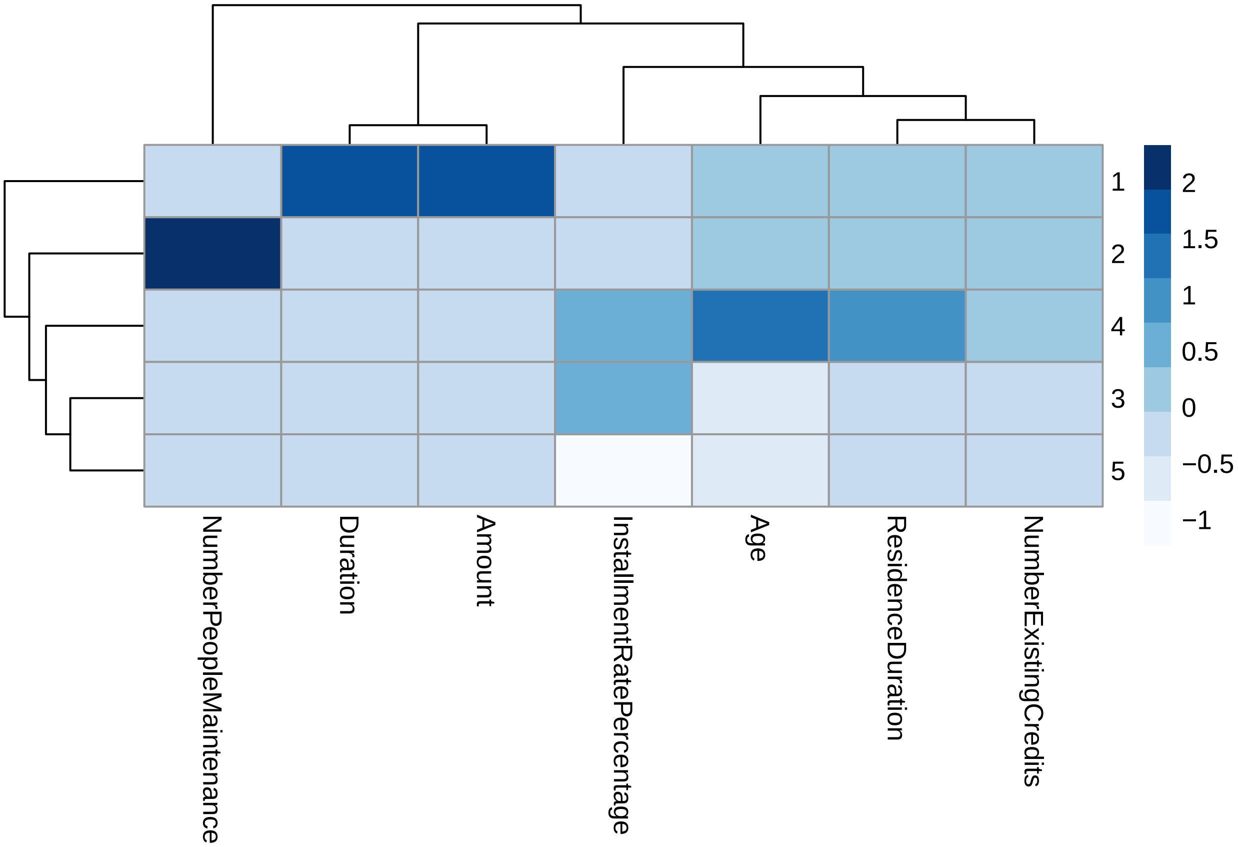
a. VAF.

Based on 4 and 5 using VAF criterion K-means cluster with 5 clusters.

b. Interpretability of the segments

I first the centers with with 5 clusters.

```
library(pheatmap)
library("RColorBrewer")
pheatmap(res_km$centers[1:5,], color=brewer.pal(9,"Blues"))
```



Based on the plot Interpretation as follow:

clusters. 1   high #People Maintenance.

clusters. 2   low age high installment rates.

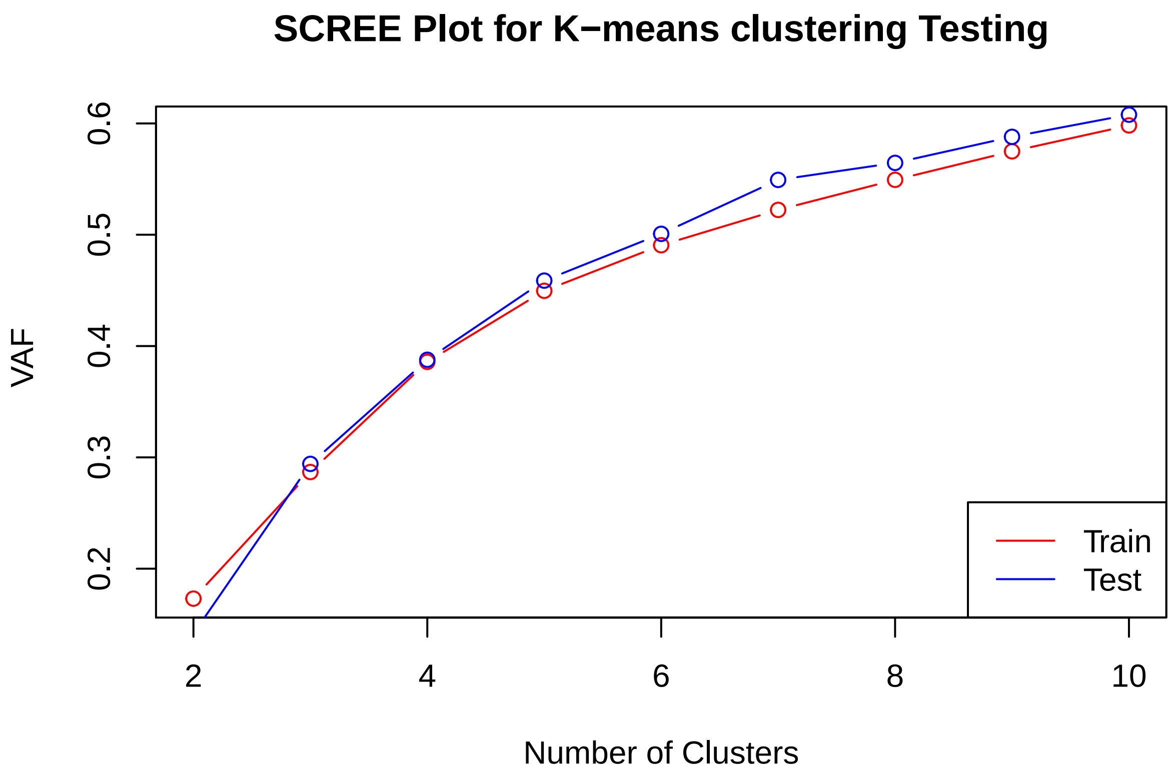
clusters. 3   high Amount and Duration.

clusters. 4   high age,Residence Duration and installment rates.

clusters. 5   low installment rates.

plot it.

```
plot(res$k,  res$VAF_train, type = 'b',col='red', xlab = "Number of Clusters",
     ylab = " VAF", main = "SCREE Plot for K-means clustering Testing")
lines(res$k,  res$VAF_hldout, type = 'b',col='blue', xlab = "Number of Clusters")
legend(x = "bottomright",
       legend = c("Train","Test"),
       lty = c(1,1),
       col = c("red","blue"))
```



By applying k means clustering the elbow appears at 5 number of clusters. Using VAF 5 cluster will be suitable here.

```
print("Clusters sizes")
```

```
## [1] "Clusters sizes"
```

```
for(i in 2:5){
  cat(paste0("Number of Cluster ",i," relative size"))
 print( round(km_hold_results[[i]]$size))
}
```

```
## Number of Cluster 2 relative size[1] 223  77
## Number of Cluster 3 relative size[1]  48 191  61
## Number of Cluster 4 relative size[1] 132  48  75  45
## Number of Cluster 5 relative size[1]  52  61 102  37  48
```

Based on relative size I think 5 clusters will be most suitable.

```
set.seed(123)

VAFS_otrain <- numeric(10)
VAFS_oholdout <- numeric(10)
kom_train_results <- list()
for(i in 3:5){
  # train
  kom_res <- komeans(standard.data[train,], nclust  = i, lnorm=2,tolerance=.00001,nloops = 100,seed=123
  kom_train_results[[i]] <- kom_res
  VAFS_otrain[[i]] <- kom_res$VAF
  # holdout
    kom_res <- komeans(standard.data[-train,], nclust = i,lnorm=2,tolerance=.00001,nloops = 100,seed=12
  VAFS_oholdout[[i]] <- kom_res$VAF
```

```r
}
res_komeans <- data.frame(k = 2:10, komeans_VAF_train = VAFS_otrain[2:10],komeans_VAF_hldout = VAFS_ohol
res_komeans[2:4,]
```

```
##   k komeans_VAF_train komeans_VAF_hldout
## 2 3         0.3485329          0.3514963
## 3 4         0.4603511          0.4778457
## 4 5         0.5921699          0.5862789
```

Print 5 clusters k-menas and komeans solutions

Print both clusters VAF

```r
cbind(res,res_komeans[,2:3])
```

```
##    k VAF_train VAF_hldout komeans_VAF_train komeans_VAF_hldout
## 1  2 0.1730995  0.1416352         0.0000000          0.0000000
## 2  3 0.2868387  0.2941430         0.3485329          0.3514963
## 3  4 0.3857962  0.3876859         0.4603511          0.4778457
## 4  5 0.4496079  0.4587734         0.5921699          0.5862789
## 5  6 0.4905909  0.5008052         0.0000000          0.0000000
## 6  7 0.5222883  0.5492414         0.0000000          0.0000000
## 7  8 0.5493096  0.5645464         0.0000000          0.0000000
## 8  9 0.5749203  0.5879431         0.0000000          0.0000000
## 9 10 0.5981831  0.6078495         0.0000000          0.0000000
```

Print member in both clusters

```r
table(km_res$cluster)
```

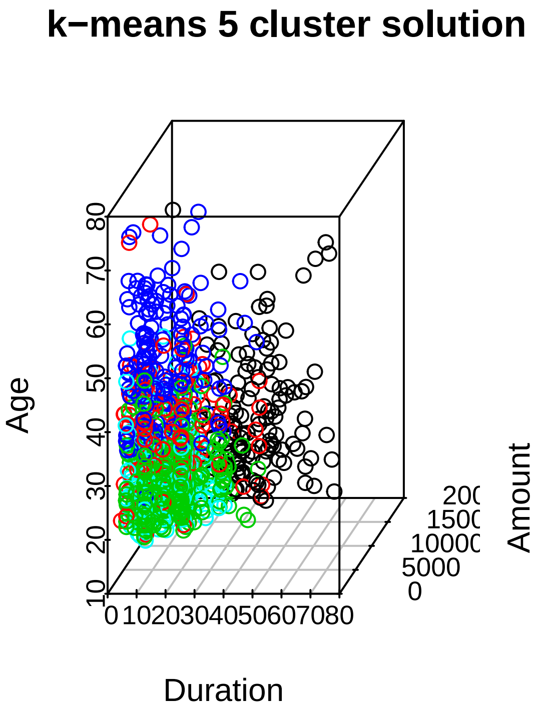
```
## 
##  1  2  3  4  5  6  7  8  9 10 
## 36 27  9 42 30 34 38 51 27  6
```

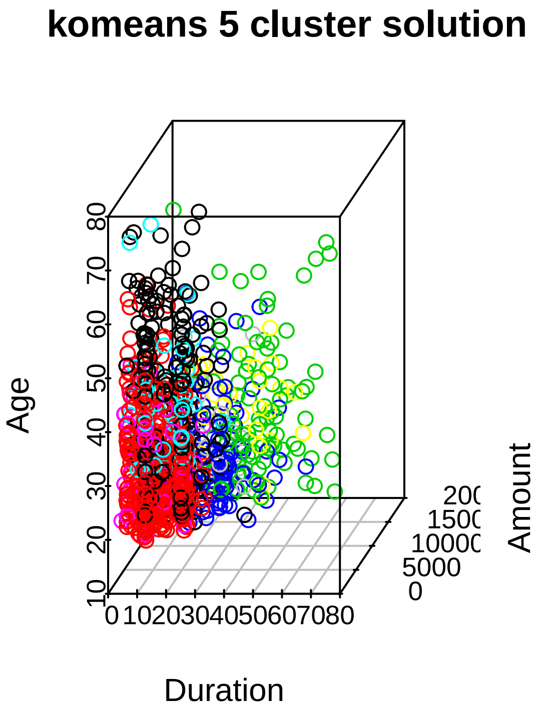
```r
table(komean_res$Group)
```

```
## 
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17 
##  51  95  38  23  20  22   7   3  46 126  24  17  18  19  12   8  41  87 
##  18  19  20  21  22  23  24  25  26  27  28  29  30  31 
##  15  22   4   7   2   3  81 113  26  40  13   9   6   2
```

Plot the clusters

```r
par(mfrow = c(1,2))
scatterplot3d(data[,1],data[,2],data[,5],xlab="Duration",ylab="Amount",zlab="Age", main= "k-means 5 clus
scatterplot3d(data[,1],data[,2],data[,5],xlab="Duration",ylab="Amount",zlab="Age",main = "komeans 5 clu
```


k−means 5 cluster solution


komeans 5 cluster solution

komeans has higher VAF than k means clustering. Based on the 3d scatter-plot it seems that   kmeans clustering is more interpretable than komeans clustering. Finally We choose k-means clustering with 5 cluster over komeans overlapping cluster solution.

By doing kmeans and komeans clustering based on VAF and Scree test I finally choose kmeans clustering with 5 cluster as final solution.

```
data$Group <- res_km$cluster
data %>%
  group_by(Group) %>%
  summarise_all(list(mean))
```

```
## # A tibble: 5 x 8
##   Group Duration Amount InstallmentRate~ ResidenceDurati~   Age
##   <int>    <dbl>  <dbl>            <dbl>            <dbl> <dbl>
## 1     1     40.8  8664.             2.63             2.82  35.7
## 2     2     17.8  2731.             2.81             2.93  38.7
## 3     3     19.0  2084.             3.78             2.44  29.4
## 4     4     16.8  2204.             3.39             3.74  50.5
## 5     5     16.2  2881.             1.60             2.67  29.9
## # ... with 2 more variables: NumberExistingCredits <dbl>,
## #   NumberPeopleMaintenance <dbl>
```

```
print("Cluser Memberships")
```

```
## [1] "Cluser Memberships"
```

```
table(res_km$cluster)
```

```
##
##   1   2   3   4   5
## 139 139 337 183 202
```

1   2   3   4   5

```
139 337 139 183 202
```

Cluster 1 has 139 members with Highest #People Maintenance

Cluster 2 has 337 members with lowest age second highest Interest Rate.

Cluster 3 has 139 members with highest Amount and duration

Cluster 4 has 183 members with highest age and Residence Duration.

Cluster 5 has 202 members with the lowest installment rates.

We'll randomly choose 30 people from each cluster and recruit them over telephone.

We'll try to recruit consumers from diverse background to make the sample unbiased for each segment.

We'll ask people about the 7 variables used in the clustering. Based on the answer we'll assign people closer to the cluster centers.

We can use principle component analysis for column reduction.

1. First take the data .

2. Select appropriate number of principal components for columns

3. Do principal component analysis on the columns

4. Extract the Principal Components.

5. Do clustering on the principal components.